L Data presentation and interpretation

L2	Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population (calculations involving regression lines are excluded). Understand informal interpretation of correlation. Understand that correlation does not imply causation.
L4	Recognise and interpret possible outliers in data sets and statistical diagrams. Select or critique data presentation techniques in the context of a statistical problem. Be able to clean data, including dealing with missing data, errors and outliers.

L Data presentation and interpretation

Students should be able to:

- interpret a scatter diagram, to include visual recognition of outliers (as stated in section L4)
- recognise and name positive, negative or no correlation as types of correlation
- understand that, in this qualification, we consider only linear correlation, but that other types are possible. (Link here to F6, where exponential and power laws are considered.)
- recognise and name strong, moderate or weak correlation as strengths of correlation in cases where comparison is possible
- understand that correlation does not imply causality
- state and use the fact that $-1 \leq r \leq 1$
- interpret, in context, correlation by considering a scatter diagram or a given value of r
- appreciate that interpretation of scatter diagrams by eye is not generally reliable
- interpret the gradient and intercept of a regression line in context.

Note: students will **not** be required to calculate a product moment correlation coefficient. Students will **not** be required to calculate the equation of a regression line or to make predictions using calculations based on the equation of a regression line.

Students will not be expected to plot scatter diagrams, but should be encouraged to use software to plot data from the LDS on scatter diagrams.

L Data presentation and interpretation

Students should be able to:

- identify outliers either from a given rule or from observation of a given diagram
- comment on the likely effect of removing the outlier
- identify clear errors in data and comment on or suggest subsequent actions needed
- select which of the representations in sections L1 and L2 is appropriate for representing given data sets
- criticise, in context, a given representation.

9.4 Bivariate Data

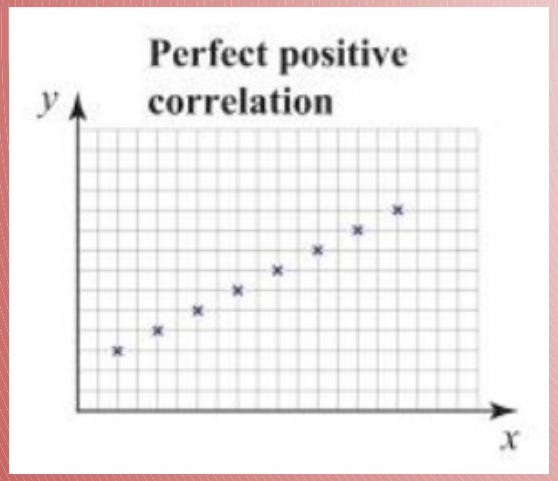
- Data relating to pairs of variables is called **bivariate** data.
- Variables that are statistically related are said to be **correlated**.
- There are three types of correlation: positive, negative and zero.
- If the variables increase together, they have **positive** correlation.
- If one variable increases as the other decreases, they have **negative** correlation.
- Two variables can also be uncorrelated and the data is then said to have **zero** correlation.

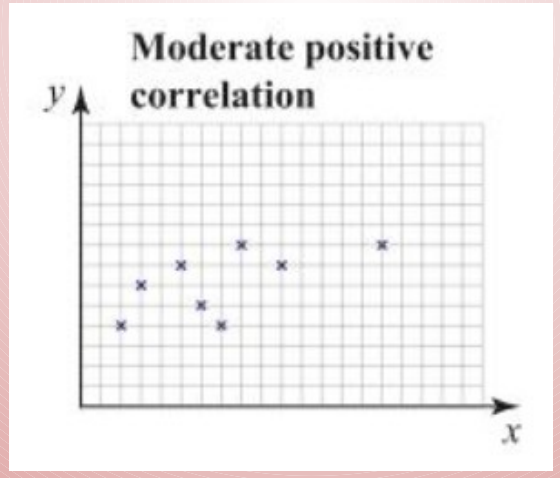
9.4 Bivariate Data

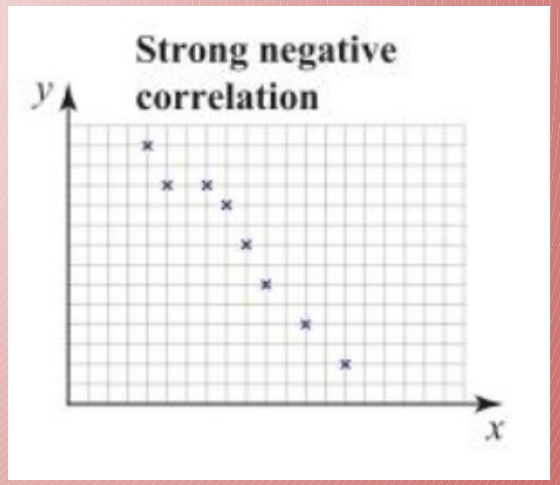
- Correlation can be identified from a **scatter diagram**.
- A scatter diagram shows both the **type** and the **strength** of the relationship between two variables.
- The **independent** variable (or explanatory variable) is the variable you can directly control. This goes on the horizontal axis.
- The **dependent** variable (or response variable) is the variable you think is being affected. This goes on the vertical axis.
- If two variables are correlated you can draw a ‘line of best fit’ (called the **regression line**).

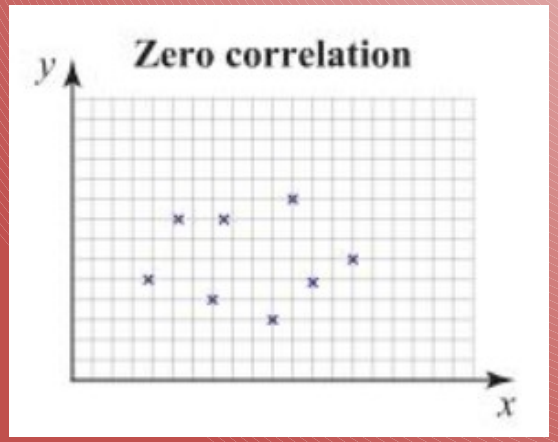
9.4 Bivariate Data

Types and strength of correlation









There is no exact definition for strong, moderate or weak correlation so it is useful to quantify this numerically (the product moment correlation coefficient,)

Perfect positive correlation,

Perfect negative correlation,

9.4 Bivariate Data

Product-Moment Correlation Coefficient ()

Strength of Correlation **Range:**

r	Description
	“weak” or “very weak”
or	“moderate”
or	“strong”
above or below	“very strong”

9.4 Bivariate Data

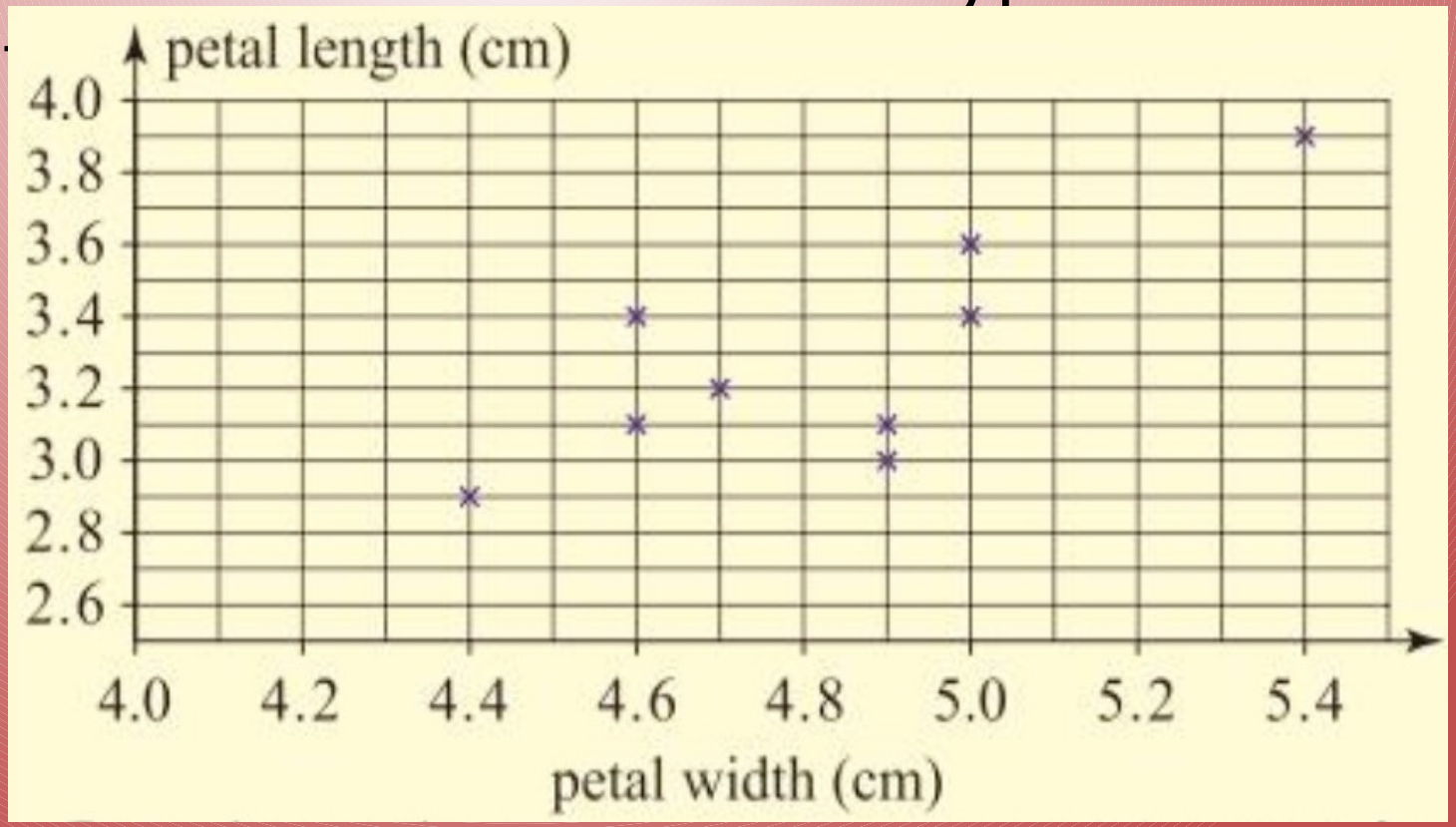
Describing Correlation

- Give the type and strength.
- Contextualise this to the question, e.g. the higher the ..., the more the ...

9.4 Bivariate Data

Example 1

The scatter graph shows widths and lengths of the petals of 9 roses. Describe the type and strength o



This relationship shows **moderate positive correlation**

9.4 Bivariate Data

The Regression Line

- The regression line is a line of best fit.
- The regression line of y on x is written in the form: .
- You can use a calculator to calculate this, however you will not be expected to do this in the exam.
- You may be given the equation and asked to interpret this.
- The coefficient tells you the change in for each unit change in . If then there is positive correlation, if then this shows

9.4 Bivariate Data

Estimating Using the Regression Line

We can predict values of the response variable (dependent variable) from the regression line (LOBF) in two ways:

1. **Interpolation** – predict values of the response variable for values of the explanatory variable that are within the range of collected data.

   This is **reliable** since the observed data supports the prediction.

9.4 Bivariate Data

Estimating Using the Regression Line

We can predict values of the response variable (dependent variable) from the regression line (LOBF) in two ways:

2. **Extrapolation** – predict values of the response variable for values of the explanatory variable that are outside the range of collected data.

This is much more **unreliable** as the data only provides evidence that the regression line is accurate for values within the

9.4 Bivariate Data

Example 2 – page 252

The following data gives an ice-cream parlour's daily sales figures, y, in hundreds of pounds, against the number of hours sunshine, x, on six consecutive Saturdays

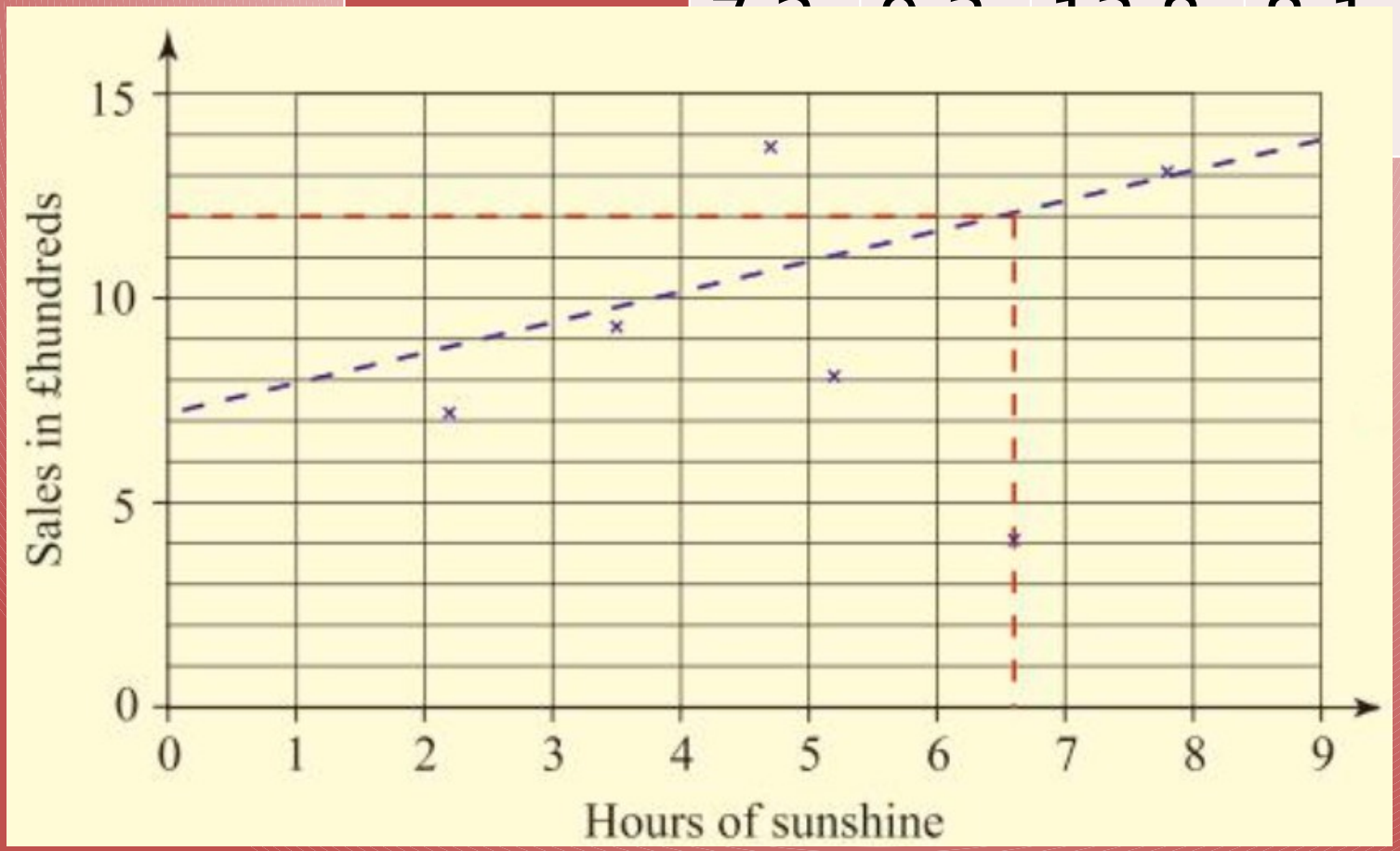
x, hours	2.2	3.5	4.7	5.2	6.6	7.8
y, £100s	7.2	9.3	13.8	8.1	4.1	13.1

a) Plot a scatter diagram to represent this data and write down the type and strength of correlation

9.4 Bivariate Data

Example

x, hours	2.2	3.5	4.7	5.2	6.6	7.8
					4.1	13.1

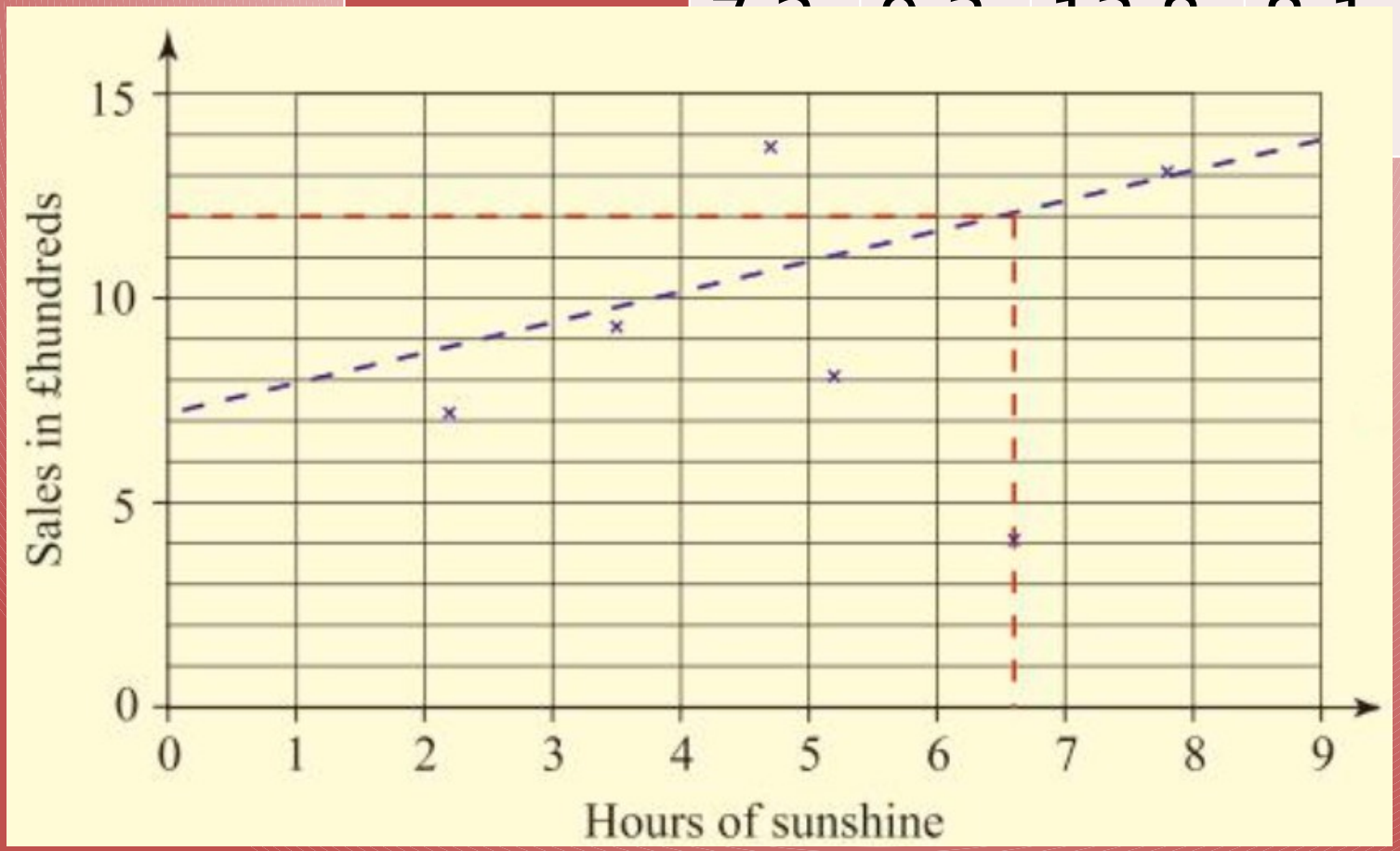


The scatter diagram shows moderate positive correlation.

9.4 Bivariate Data

Example

x, hours	2.2	3.5	4.7	5.2	6.6	7.8
					4.1	13.1

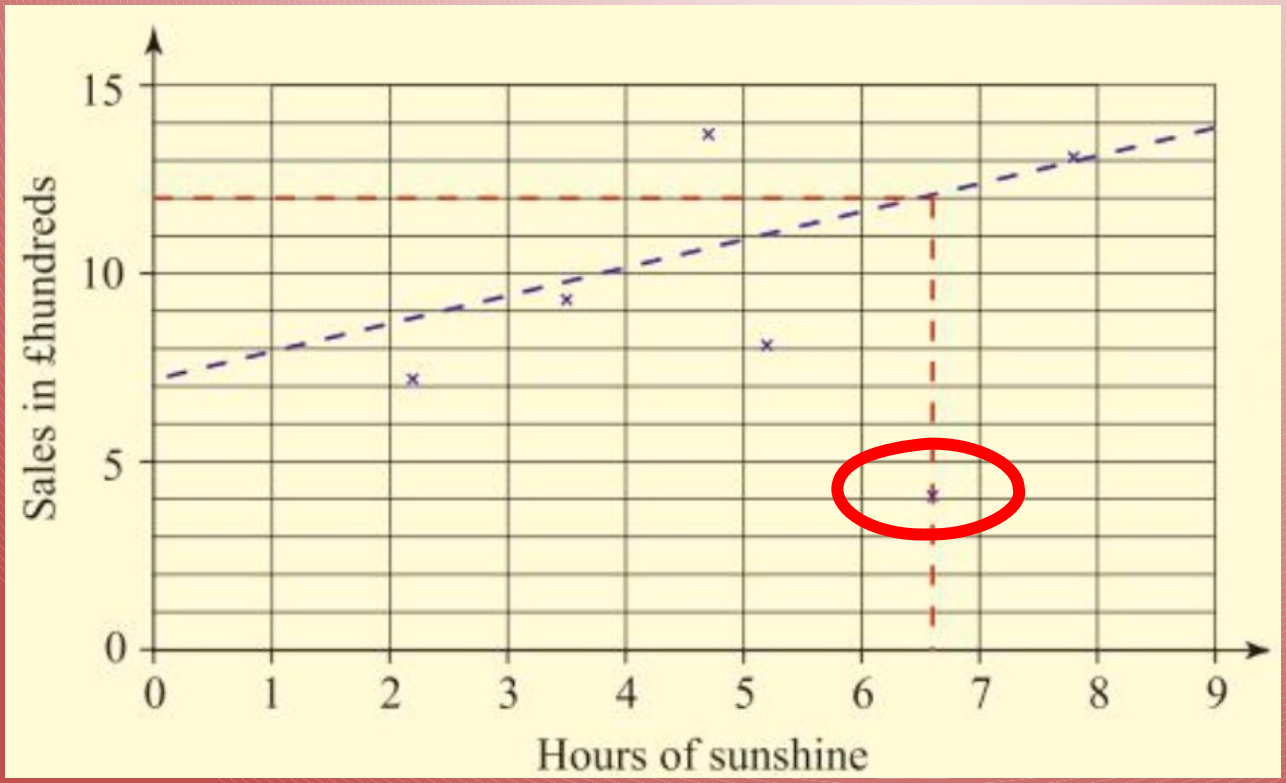


You can use menu 6, option 2 to work out the value.
Type the data in and use OPTN, regression calc.
This also gives the equation of the

9.4 Bivariate Data

Example 2 – page 252

b) One of the points relates to a day when there was a power cut and the ice-cream parlour was closed for several hours. Write the most likely



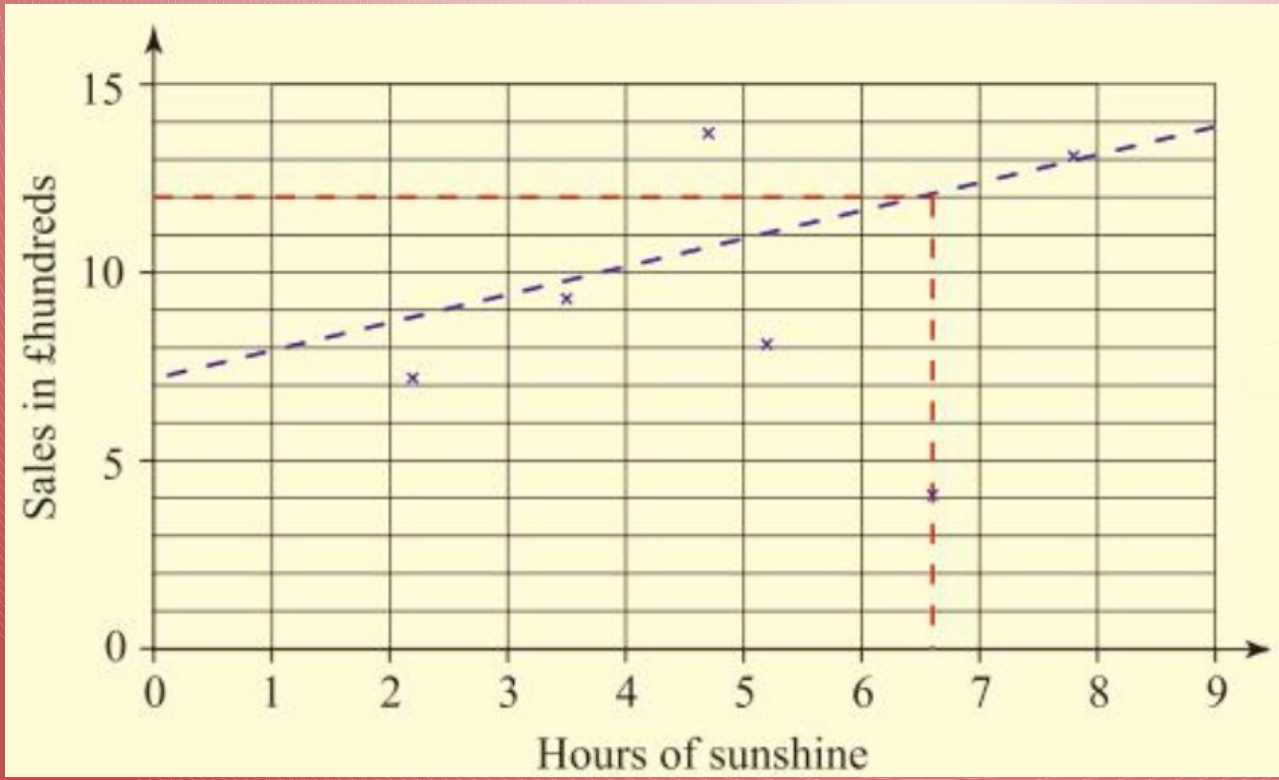
(6.6, 4.1)

Note that we could recalculate after removing this point…

9.4 Bivariate Data

Example 2 – page 252

c) For the data identified in part b, give a sales figure that could be expected if the power cut



Draw a line of best fit, ignoring the point in part b, and use this to predict sales with 6.6 hours sunshine:

9.4 Bivariate Data

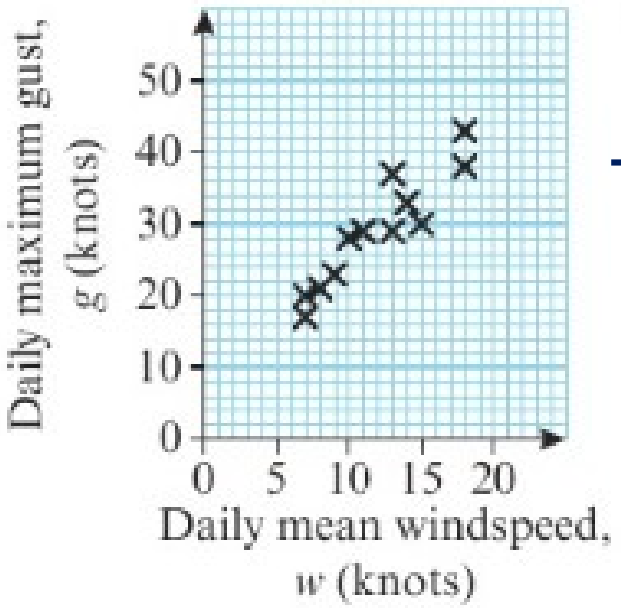
Example 3a

Note this is the Edexcel LDS!

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data was plotted on a scatter diagram:



a Describe the correlation between daily mean windspeed and daily maximum gust.

There is a strong positive correlatio

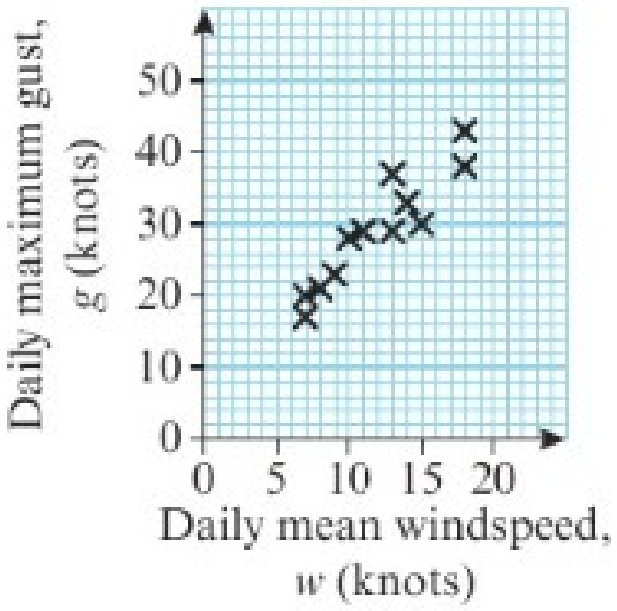
9.4 Bivariate Data

Example 3b

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data was plotted on a scatter diagram:



The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$.

b Give an interpretation of the value of the gradient of this regression line.

With each 1 knot increase in daily mean windspeed, the daily maximum gust increases by 1.82 knots.

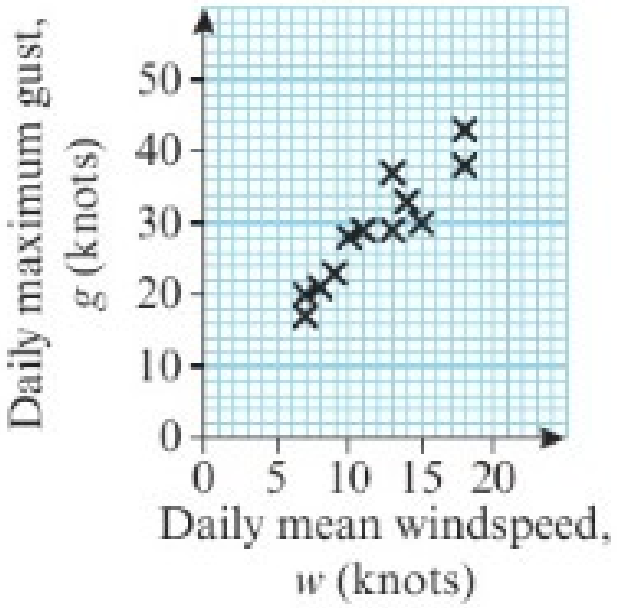
9.4 Bivariate Data

Example 3c Note – you can have non-linear correlat

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data was plotted on a scatter diagram:



The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$.

c Justify the use of a linear regression line in this instance.

The correlation suggests that there is a linear relationship between and so a linear regression line is a suitable

9.4 Bivariate Data

Example 4: from the Large Data Set

Do cars with larger engines emit more CO_2?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Reference Number	Make	Propulsion TypeId	BodyType Id	GovRegion	Keeper Titlel	EngineSize	YearRegistered	Mass	CO2	CO	NOX	part	hc
2	440	VAUXHALL	1	96	London	1	1598	2002	1970	190	0.219	0.026		0.037
3	1465	VAUXHALL	1	14	South West	5	1398	2016	1163	118	0.463	0.01		0.031
4	3434	VOLKSWAGEN	1	14	South West	2	1395	2016	1316	113	0.242	0.033		0.048
5	1801	VAUXHALL	1	14	South West	4	1598	2016	1355	159	0.809	0.012		0.051
6	2330	BMW	2	13	South West	5	1995	2016	1445	114	0.18	0.023		
7	2216	FORD	2	6	South West	5	1499	2016	1425	98	0.354	0.074		
8	1323	VAUXHALL	1	14	South West	5	1398	2016	1163	118	0.463	0.01		0.031
9	1396	VAUXHALL	1	14	South West	5	1398	2016	1163	118	0.463	0.01		0.031
10	264	TOYOTA	1	14	London	4	1998	2002	1440	212	0.76	0.04		0.08

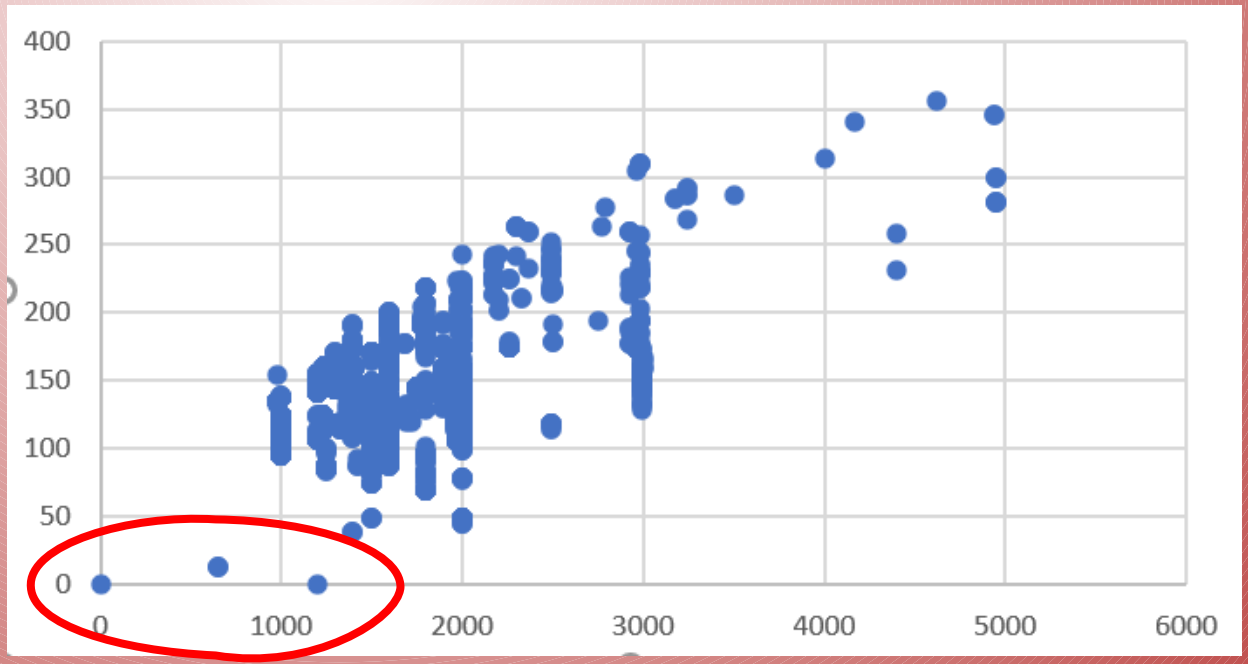
en up AQA-AS-A-MATHS-LDS-2019-2020.xls
m the “Car Data” tab, highlight columns G and J.
m the “Insert” menu, select “Scatter”.

9.4 Bivariate Data

Example 4: from the Large Data Set

Do cars with larger engines emit more CO_2?

Engine size should appear on the -axis, CO_2 emissions will be on the -axis.



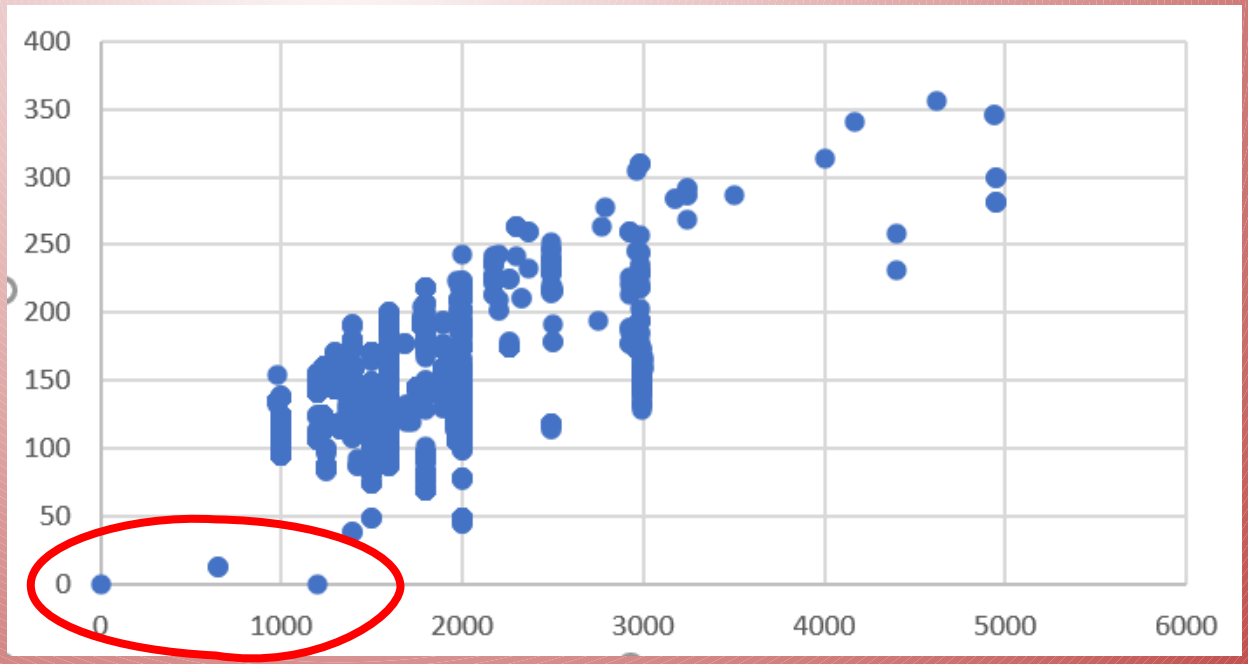
It would be worth investigating these cars to see whether there are any errors in the data.

9.4 Bivariate Data

Example 4: from the Large Data Set

Do cars with larger engines emit more CO_2?

- Can a car have an engine size of zero?
- What sort of cars have very low (or zero) emissions?



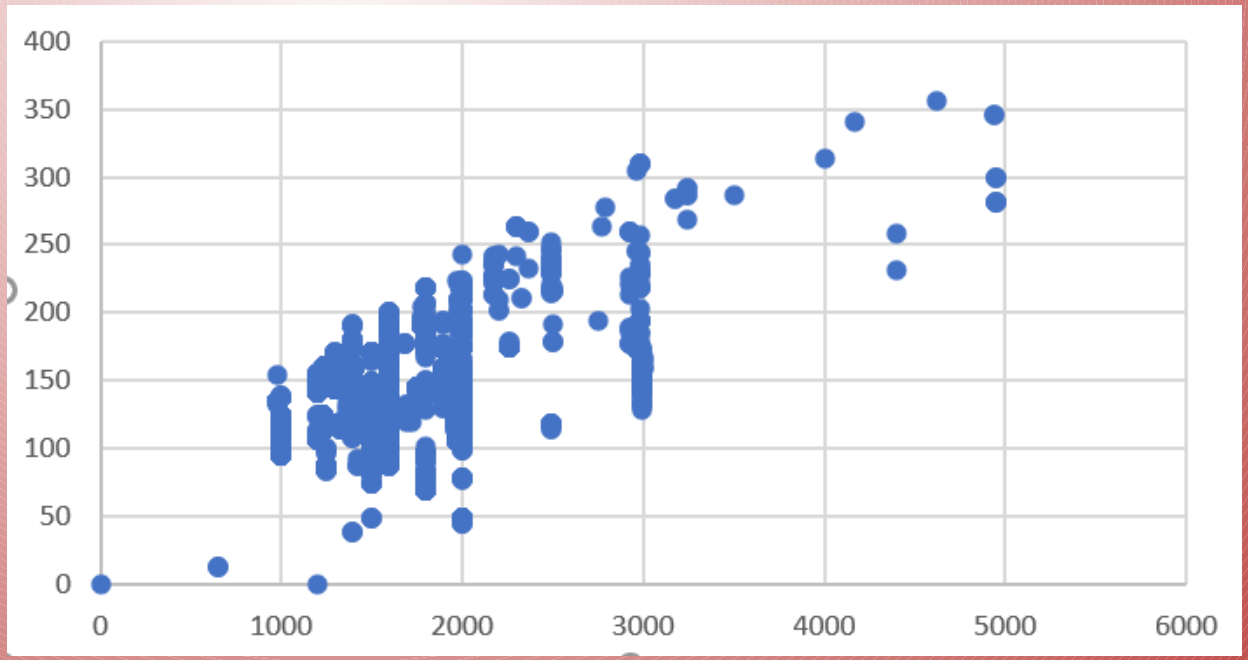
Have a look at the PropulsionTypeId column. You can find what the numbers mean by looking on the "Field Values" tab.

9.4 Bivariate Data

Example 4: from the Large Data Set

Do cars with larger engines emit more CO_2?

What type of correlation does this show?



We can find the correlation coefficient as follows:

Type "=correl(G2:G11,J2:J11)" into an empty cell.

9.4 Bivariate Data

Correlation and Causation

Just because two variables correlate, we should not assume that changes in one variable are **causing** changes in the other.

For example, rates of diabetes and annual income correlate for certain groups, but this is because they both relate to dietary intake.

When a change in one variable *does* affect the other, they have a **causal connection**. Correlation without a causal connection is known as a **spurious correlation**.

9.4 Bivariate Data

Correlation and Causation

Does 5G **cause** coronavirus?

9.4 Bivariate Data

Correlation and Causation

Some spurious correlations:

https://www.tylervigen.com/spurious-correlations

Number of people who drowned by falling into a pool correlates with the number of films that Nicholas Cage starred in:

Per capita cheese consumption correlates with the number of people that died by becoming tangled in